DOCUMENT RESUME

ED 480 059 CG 032 632

| | |
|---|---|
| AUTHOR | Behuniak, Peter |
| TITLE | Education Assessment in an Era of Accountability. |
| PUB DATE | 2003-08-00 |
| NOTE | 15p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608. |
| PUB TYPE | Information Analyses (070) |
| EDRS PRICE | EDRS Price MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Accountability; *Achievement Tests; *Educational Assessment; Educational Change; Educational Testing; Public Education; *Test Use; *Trend Analysis |

ABSTRACT

Two trends have converged during the past three decades to change the face of public school education in America. First, achievement testing has been greatly expanded in terms of both the quantity of tests available and the number of uses for the information collected from testing. Second, there has been a significant increase in the development of accountability systems for the purpose of fostering educational reform. This chapter discusses these developments in three sections. The first section describes some of the key influences and history behind these trends. The second section examines how the widespread adoption of accountability systems is affecting the types of achievement tests being created, the frequency of their use, and the purposes to which their results are applied. The third section focuses on a number of areas related to these trends that are of particular concern to educators. (Contains 16 references.) (GCP)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 480 059

# *Education Assessment in an Era of Accountability*

By
Peter Behuniak

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CG032632

BEST COPY AVAILABLE

# Chapter 24
# Educational Assessment in an Era of Accountability

*Peter Behuniak*

Two trends have converged during the past three decades to change the face of public school education in America. First, achievement testing has been greatly expanded in terms of both the quantity of tests available and the number of uses for the information collected from testing. Second, there has been a significant increase in the development of accountability systems for the purpose of fostering educational reform. Although there are many different ways of designing a system of accountability, virtually all approaches employ achievement test results as one, usually a central, component. As a result, the increases in student testing and greater demands for accountability have interacted to make learning and teaching in U.S. public schools at the beginning of the twenty-first century quite different than they were prior to the 1970s.

In this chapter I discuss these developments in three sections. The first section describes some of the key influences and history behind these trends. The second section examines how the widespread adoption of accountability systems is affecting the types of achievement tests being created, the frequency of their use, and the purposes to which their results are applied. The third section focuses on a number of areas related to these trends that are of particular concern to educators. I offer suggestions to illustrate how teachers and others involved in public education can effectively respond in the current environment.

## How Did We Get Here?

Once upon a time the term *accountability* was nearly synonymous with *responsibility.* Students were responsible for learning their lessons. Teachers were responsible for presenting important topics in class and helping students as needed. Administrators were responsible for supporting teachers, monitoring their effectiveness, and communicating with parents. Parents had the responsibility of ensuring their children

were good students. Similarly, district administrators were accountable to local boards of education, the local boards were accountable to state agencies and the voters, and so on. In short, everyone was accountable to someone.

Of course, not everyone accepted their role, nor were they equally effective even if they did accept it, but at least the system was easy to understand. Everyone was expected to display responsible behavior (e.g., learning, teaching, parenting) to the satisfaction of at least one other individual or group. If someone did not do his or her job, the next link in the chain would do the responsible thing and intercede. At least that was the plan.

This shared system of responsibility does not place a very high demand on achievement testing. Through most of the early and middle years of the twentieth century, achievement tests served primarily to provide teachers with instructional feedback and confirmation of student learning. In some cases, results were shared with parents or summarized for administrative purposes. In other cases, the information remained with the teacher. Testing was all handled in a low-key manner.

This began to change in the 1970s. Legislatures and educational bureaucracies, particularly at the state level, discovered that standardized achievement tests could be pressed into service as instruments of reform. Throughout the country, minimum competency testing (MCT) was introduced to the public schools. This was a new kind of testing in which each student's performance would be judged against a previously established standard (the minimum competency standard) to determine whether adequate learning was occurring. Some MCT programs were designed to focus attention on the teachers by calling for improved teaching strategies when test results were judged to be too low. Other MCT programs held the students responsible by applying sanctions such as the denial of a promotion or a diploma. Some MCT models tried to hold educators, students, and parents responsible. All of these programs, however, demonstrated clearly that the race to high-stakes testing had begun.

This top-down approach of using educational tests as a hammer to force change became a source of concern almost as soon as it was introduced. Jaeger & Tittle (1980) worried in the prologue to their book *Minimum Competency Achievement Testing* that the implementation of MCT programs was moving ahead too quickly without adequate attention to its consequences. In words that now seem prophetic, they wrote, "Comparatively little attention has been directed to such larger issues as the need for minimum competency testing, the problems it

seeks to solve, its likely effects on the structure and operation of the schools, and its consequences for those directly involved in elementary and secondary education, as well as for our larger society" (p. vii).

Eventually, after about a decade, the popularity of MCT waned as some of its negative effects were realized, such as narrowed or watered-down curriculum and reduced student motivation. In its place, new testing programs were implemented with higher standards and broader content. Reports such as *A Nation at Risk* (National Commission on Excellence in Education, 1983) fueled national concerns about the effectiveness of U.S. schools. The old model of shared responsibility was gradually replaced with calls for educational reform and more formal systems of accountability. The political lessons learned years earlier dominated the landscape. The country turned its attention to school and district accountability, then to standards-based accountability (Linn, 1998). The tests mandated in the 1980s and 1990s were implemented with even higher stakes, including programs where the allocation of financial and other resources, the security of teachers' jobs, and even the continued existence of specific schools rested on the results of standardized tests. It suddenly became crucial to have high-quality tests in place, given how much was depending upon them. As Paul E. Barton (1999, p. 6) stated, "Improving testing is important because testing has become, over the last 25 years, the approach of first resort of policymakers."

It is worth noting that the expanded role of testing in the public schools cannot be dismissed as merely a political gimmick. The expansion of testing and accountability systems has support that extends well beyond state legislatures and education departments. Rose and Gallup (2001) report that 66 percent of the U.S. public believes that the emphasis on achievement testing in public schools is at the right level or should be increased. Interestingly, this support climbs to 73 percent when the parents of public school students are polled. Three quarters of the public indicated they support President Bush's proposal to hold the public schools accountable for how much students learn. Phelps (1998) considered a large number of surveys and polls and concluded, "The general public, parents, students, and often teachers want more testing, and they want higher-stakes testing. Perhaps they do because they are not looking at testing's problems out of context, in isolation from consideration of the real alternatives to testing, as testing's critics often are. They are considering testing against the alternatives, and they think that some testing, more testing, is better" (p. 16). Although some would take exception to this claim, particularly regarding the

support from teachers (see, e.g., Wassermann, 2001), it is evident that the trend toward more testing and more high-stakes testing has widespread support.

## Where Are We Now?

The increased use of tests for the purpose of holding schools accountable has caused much debate and discussion. Warnings have been issued, much like the one Jaeger and Tittle sounded more than two decades ago, that point out the many potential negative effects of such high-stakes applications of tests (Popham, 1999; Shepard, 2000). Concerns include a narrowing of the curriculum, corruption of sound teaching strategies, lessening of attention in classrooms on higher-order skills, lowering of student and teacher motivation, and reduction of attention to students' individual needs. Some researchers, however, have offered strategies for dealing with the current environment (Gallagher, 2000; McColskey & McMunn, 2000).

One positive result of the focus on testing for accountability purposes has been a greater effort to produce tests of higher quality. This attention to test design often grew directly from criticism of the shortcomings or limits of available tests. For example, it is much more common today for achievement tests to include varied formats, with students explaining their work and completing extended performance tasks in addition to answering short-answer or selected-response questions. This has improved the capacity of tests to measure a broader and deeper range of student achievement.

The creation of content standards and the design of achievement tests consistent with those standards is another important development in the evolution of high-stakes testing. It was not uncommon 15 or 20 years ago to create and administer an assessment first, then worry about sharing descriptions of what the test measured later. The increased use of tests in high-stakes situations has made this a less frequent occurrence. Attention is now given to aligning the material covered on a test with established content standards and to publicizing those standards well before the first test administration.

Another area of improvement involves the use of technology to enhance how test results are shared and the speed of returning those results to schools. Higher stakes mean higher interest levels in the test results. Many accountability programs now provide customized reports tailored to the test users' needs, web-based access to data, easy-to-use software for examining the results, and a variety of CD- or web-based

tutorials to improve educators' understanding of the test results. Though the volume and complexity of tests have generally increased, many programs have succeeded in maintaining or reducing the time between the test administration and the reporting of the results. One of the more promising developments in the past few years has been advances in making computer-based testing suitable for certain uses with large-scale assessments (Bennett, 2001).

Yet, despite all the improvements in test design and implementation, concerns persist regarding the wisdom of depending too much on one or a few assessments. Even educators and assessment specialists who applaud the improvements in the quality of tests voice doubts about whether using them in accountability systems will have a positive effect on schools (Hilliard, 2000). Haertel (1999), for example, acknowledges the benefits of using performance assessments but questions the underlying assumptions of test-centered accountability. He concludes, "Regardless of the value of performance assessments in the classroom, a measurement-driven reform strategy that relies on performance assessment to drive curriculum and instruction seems bound to fail" (p. 666).

The title of this section is posed as a question: "Where are we now?" The answer to this question will be somewhat different for each classroom in America. One of the impediments to a meaningful discussion of high-stakes testing is that the actual effect of a system of testing and accountability on any particular student, teacher, or school depends upon many components of the system in question and how those components interact. In a review of the assessments and accountability systems planned or in place in each of the 50 states, Linn (2001) found them to differ on multiple dimensions, making the evaluation or categorization of the systems difficult. Consider, for example, one component common to many state accountability systems: Students are required to pass a test to earn a high school diploma. Any two states that have such a requirement may differ on the content areas tested, the rigor of the standards, the number of times a student may retake the test, and the accommodations offered to some or all students. In addition, two students attending different schools (or having different teachers in the same school) may receive instruction that varies in its focus on the content covered by the test.

This means that if 1,000 schools operate under a statewide program of testing and accountability, there are potentially 1,000 combinations of factors producing a system of accountability that is unique to each school. If achievement testing is to play a positive role in improving

education, all the stakeholders in a given community will need to examine critically the components of the system operating in their own backyard. Some aspects will be found useful and productive. Others may be ineffective or even counterproductive and should be reconsidered. Overall, it is important to realize that there is no one best model and that many local factors may affect the way the system operates.

## Issues and Strategies

One of the few principles that virtually every policymaker and stakeholder involved in education agrees on is the central importance of teachers in any reform effort. There is more than a little irony in the fact that an era of mandated testing and stringent accountability systems could have the unintended effect of disenfranchising the very individuals crucial to the public schools' mission. There are, however, strategies and actions that can be useful in ensuring that educators' voices are heard amid all the cries for reform.

The issues and strategies identified in this section focus on basic principles of assessment and instruction and how these elements interact. They are offered to serve dual purposes. First, they highlight key elements of educational assessments that require scrutiny to ensure that sound tests and testing practices are in place. This is a type of watchdog function, and no one is better positioned to fulfill this function than the individuals who regularly administer and proctor the tests, report the test results to students and parents, and interpret the implications of those results. The second purpose is more oriented to professional development. Stated simply, educators who are more knowledgeable about the form and function of the tools of their trade will be in a stronger position to express their views and concerns effectively.

### *Learn Basic Measurement Principles*

Assessments are tools of the profession of education. High-stakes tests, low-stakes tests, selected-response formats, performance assessments, commercially available standardized batteries, and exit exams are all just variations that may be more or less appropriate for any given purpose. In many cases, one test used for a certain purpose will have both positive and negative consequences. It is important that the educators involved in using these tests have a solid fundamental understanding of the measurement principles on which these tests are based.

This does not mean that every teacher must become a measurement expert. Psychometricians and assessment specialists have a productive role to fill in the field of education, just as do specialists in reading, music, or administration. One course taken during an undergraduate teacher training program is probably not sufficient to provide educators with a working knowledge of measurement principles such as reliability, validity, bias, errors of measurement, and test standardization. They can acquire more information about these principles through such activities as continuing education coursework, private study, and in-service professional development. Regardless of an individual educator's disposition toward educational testing in general or toward any specific test, gaining a better understanding of the principles that guide their development will enhance his or her chances of taking full advantage of reasonable test applications and provide a credible basis to support criticism of unwarranted applications.

*Know Each Locally Administered Test*

If educational tests are to function as effective tools for guiding students' learning, educators must be prepared to select the right tool for the job. The first step in this process is to become familiar with all the tests being used with local students. This includes all assessments in use regardless of whether they are optional, mandated locally, or mandated externally. Increasing familiarity with the assessments in use could begin with background information, such as who developed the test, for which ages or grades it was designed, and whether evidence of technical quality has been provided. Although some educators are not experienced enough to judge the technical merits of a particular assessment, all teachers and administrators are capable of at least verifying that someone with technical skills has reviewed the tests.

Classroom assessments and teacher-made tests should also be considered. These tests are much less formal and usually do not have evidence of technical quality available. This is acceptable because of the low-key way in which they are typically used. They are important to consider, however, because students often spend more time taking many of these brief, informal tests than they do taking the higher-stakes, more formal assessments. It is not necessary for classroom assessments always to measure the same skills that the more formal assessments measure, for a teacher may well wish to use a classroom test to check on students' understanding of skills prerequisite to or otherwise separate from the content represented on other assessments. It is necessary to ensure, however, that the ways in which students are tested on multiple

assessments are not inconsistent or contradictory. For example, if students are expected to produce writing samples as part of their assessments, it could be unintentionally confusing to the students if the criteria for grading the essays differ from one test to another (e.g., spelling counts in one test but not in another).

*Consider the Purpose of Each Test*

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), the primary reference in the field, makes clear that validity depends on the purpose or purposes for which any given test is used. In fact, the very first standard (Standard 1.1) requires that every use to which a test is put be supported by reasonable evidence. In some cases, identifying the intended purpose of an assessment is easy. Any state or federal agency mandating an assessment will specify at least one intended purpose. In these cases, the judgment necessary in each school or classroom has more to do with the appropriateness of the specified purpose for the students involved. For example, consider a statewide minimum competency test intended to identify students who are not proficient in reading in order to have them receive remediation. The appropriateness of the test for students in a limited English proficient class may be questionable even if the test is reasonable for use with most students.

Sometimes an assessment has no compelling purpose. It is surprising how often an assessment that once served a useful role continues to be administered annually or periodically long after it has ceased to fulfill that purpose. An example of this might occur if a district continues to administer in elementary grades a battery of achievement tests that had been instituted years before the state agency established statewide content standards for the same grades and mandated statewide assessments aligned to those standards. The main purpose for the district assessment may no longer apply if the state assessment is filling that role. The judgment in this case turns on whether the district assessment serves any other suitable purpose or whether the students' and teachers' time would better be spent on other activities.

An additional point concerns the need to consider the technical adequacy of a test in relation to its intended use. Teachers can assist in the process of identifying any shortcomings of a test, particularly a high-stakes measure, by carefully considering test results for the students in their classes. Instances in which the results are inconsistent with existing information about the students should be questioned. All tests include error, so not every student will score exactly as other factors

might have predicted. At the same time, unusual results for large numbers of students (or extremely unusual results for one or a few students) may indicate a problem in the test design, scoring, or reporting.

*Watch for Unintended Consequences*

Almost all policymakers and administrators influencing the course of public education intend to improve the quality of teaching and learning through their actions and decisions. Unfortunately, what is intended does not always happen. Sometimes—some people would argue all too often—testing programs produce the type of negative, unintended results discussed earlier. This is the reason teachers need to be vigilant to the problems that may occur when high-stakes tests are implemented.

This vigilance requires determination and effort on the part of educators. It is not easy to maintain a balanced perspective regarding the effects of an accountability program when part of the purpose of that program is directed at you. Yet, it is undeniable that teachers are the professionals best positioned to notice if one or many students are being negatively affected in some way. If several or all teachers in a school begin to share the same observations and concerns, it is worth discussing and, perhaps, attempting to minimize or eliminate the problem.

*Rely on Multiple Indicators*

Most educators do this instinctively. For these educators, this principle is merely reassurance that it is indeed appropriate and desirable to consider all available evidence about a student when interpreting a test score. Classroom performance, grades, individual learning styles, and other test results are all useful indicators to consider.

The goal should be to bring any newly available test result into the context of all that is known about each student. If the new test scores essentially confirm existing information, a teacher has one more reason to support the instructional choices he or she is making for that student. If the new test results are at odds with some of the existing information, it is appropriate to dig deeper into the reasons for the discrepancy. It is possible that the new results are somehow invalid, perhaps because problems occurred during the test administration, the student misunderstood the directions, the test was developed or scored inappropriately, or for many other reasons. It is also possible, however, that the assessment is revealing an academic weakness or other aspect of the student's understanding that had not previously been noticed.

The role of the teacher in this circumstance is that of a diagnostician, investigating all reasonable possibilities, with help from specialists when necessary, until the discrepant information can be reconciled and a suitable program of instruction determined.

*Make Testing a Positive Classroom Experience*

When athletes compete, they give their best effort to the activity. If they did not, if they made only a token effort, the event would be meaningless. Tests of student achievement are also intended to be measures of maximum effort. They are intended to monitor how well a student can do when that student does the best he or she can do. If anything interferes with the student demonstrating his or her best work, the resulting test scores will be misleading and invalid. Many teachers make a reasonable effort to motivate students appropriately. Problems can occur, however, if teachers do either too much or too little.

Excessive test preparation is probably the most common example of how a teacher can do too much. It is appropriate to give students advance notice of an upcoming test. It is also reasonable to ensure that students are familiar with the types of questions and tasks that they are likely to encounter on the test. This is the reason large-scale tests usually are preceded by short practice tests, so that the format of the test does not surprise or confuse students. Repeatedly exposing students to practice sessions involving test questions that are similar or identical to the actual questions is neither good instruction nor good test preparation, however. This problem can become even more pronounced if a school or district administrator encourages or demands such activities. Other examples of inappropriate teaching behaviors include creating excessive student anxiety by overselling the importance of the test or coaching students during the actual test administration.

There are many ways in which a teacher can do too little to promote a positive environment. One is through indifference, for example when a teacher fails to announce or discuss the test with students in advance. Even worse is the situation where a teacher is openly critical of the purpose or nature of the assessment with students. Many teachers appear to be surprisingly unaware of the powerful depressive effect their negative comments can have on their students' motivation and results. This does not mean that a teacher cannot be critical of certain aspects of an assessment or accountability system, as the next section discusses. It does mean, however, that teachers should be circumspect in how and when they express their views.

The most positive testing experiences for students occur when

teachers and administrators work together to help students place the assessments in a balanced context. Yes, the tests are important and you should do your best work. No, this test score is not the only thing that matters. Yes, you will have a chance to practice a few questions like the ones that will be on the test. No, we will not shut down the school for a month prior to the test just to drill test questions. The key here is to prepare students effectively to demonstrate to the very highest level possible what they know and can do.

*Contribute Constructively to Improved Assessment Practices*

In some ways, this suggestion is the logical extension of the points made earlier. Take some time to learn about the principles of good testing practices and the specifics of tests to which your students will be exposed. Prepare your students for the assessment but do not overdo it. Make the most of the test results but interpret those results in relation to all other available information. Be on the lookout for unintended negative consequences on the curriculum or the students. In short, become proactive in a balanced way, acknowledging the productive and useful role that assessment can play while working to change problematic aspects of the system.

Large-scale assessments are complex undertakings. Implementing them as part of an accountability system only increases the complexity and the potential for problems. In order for these systems to function in a manner that improves public education, it is essential that all educators, including classroom teachers, contribute their varied perspectives and talents to improve them. They are the tools of our profession. It is our collective responsibility to see that they are used wisely.

## References

AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barton, P. E. (1999). *Too much testing of the wrong kind, too little of the right kind in K–12 education.* Princeton, NJ: Educational Testing Service.

✦Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives, 9,* 1–24.

Gallagher, C. (2000). A seat at the table: Teachers reclaiming assessment through rethinking accountability. *Phi Delta Kappan,* 81, 502–507.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan,* 80, 662–666.

Hilliard, A. G. (2000). Excellence in education versus high-stakes standardized testing. *Journal of Teacher Education,* 51, 293–304. (ERIC Document Reproduction Service No. EJ 613877).

Jaeger, R. M., & Tittle, C. K. (1980). *Minimum competency achievement testing.* Berkeley, CA. McCutchan Publishing.

✦Linn, R. L. (1998). *Assessments and accountability* (Technical Report No. 490). Los Angeles: Center for the Study of Evaluation, CRESST/ UCLA. (ERIC Document Reproduction Service No. ED 443865).

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (Technical Report No. 539). Los Angeles: Center for the Study of Evaluation, CRESST/UCLA.

McColskey, W., & McMunn, N. (2000). Strategies for dealing with high stakes tests. *Phi Delta Kappan,* 82, 115–120.

National Commission on Excellence in Education. (1983, April). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Department of Education.

Phelps, R. P. (1998). The demand for standardized student testing. *Educational Measurement: Issues and Practice,* 17, 5–23.

Popham, W. J. (1999). Where large scale assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice,* 18, 13–17.

Rose, L. C., & Gallup, A. M. (2001). The 33rd annual Phi Delta Kappan/ Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan,* 83, 41–58.

▼Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher,* 29, 4–14.

Wasserman, S. (2001). Quantum theory, the uncertainty principle, and the alchemy of standardized testing. *Phi Delta Kappan, 83,* 28–40.

♦ Document is included in the Anthology of Assessment Resources CD
▼ Document is available on a website

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

# NOTICE

## Reproduction Basis

| | |
|---|---|
| | This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form. |
| X | This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket"). |